```r
#1. Wanted to know where the working directory was:
getwd()
```

```
## [1] "C:/Users/Nupur Shrinet/Documents/predictive analytics/Project"
```

```r
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.6.3
```

```r
#2. After loading the cleaned data we understand that there are 1000 observationbs from 39 variables:
german<-read.csv("cleaned_data.csv")
```

```r
#3. We wanted to understand the structure of the data set by looking at the variables and their constru

str(german)
```

```
## 'data.frame':    1000 obs. of  39 variables:
##  $ Record.Id                                                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Credit.Risk                                                 : int  1 2 1 1 2 1 1 1 1 2 ...
##  $ Installment.rate_transformed                                : int  4 2 2 2 3 2 3 2 2 4 ...
##  $ Residence.Tenure_transformed                                : int  4 2 3 4 4 4 4 2 4 2 ...
##  $ Existing.credit_transformed                                 : num  2 1 1 1 2 1 1 1 1 2 ...
##  $ Dependents_transformed                                      : int  1 1 2 2 2 2 1 1 1 1 ...
##  $ Duration_transformed                                        : num  -1.26 2.315 -0.749 1.804 (
##  $ Credit.amt_transformed                                      : num  -0.788 1.063 -0.429 1.81 (
##  $ Age_transformed                                             : num  2.799 -1.2 1.199 0.844 1.5
##  $ Current.Ac.status                                           : num  1 2 0 1 1 0 0 2 0 2 ...
##  $ SavingAc.Bonds                                              : num  0 1 1 1 1 0 3 1 4 1 ...
##  $ Emp.Tenure                                                  : num  4 2 3 3 2 2 4 2 3 0 ...
##  $ Debtors.Guarantors                                          : num  0 0 0 2 0 0 0 0 0 0 ...
##  $ Housing                                                     : num  1 1 1 2 2 2 1 0 1 1 ...
##  $ Job                                                         : num  2 2 1 2 2 1 2 3 1 3 ...
##  $ Telephone                                                   : Factor w/ 2 levels "none","yes
##  $ Foreign.Worker                                              : Factor w/ 2 levels "no","yes":
##  $ credit_history_transformed                                  : num  1 2 1 2 0 2 2 2 2 1 ...
##  $ Status...Sex_female...divorce.seperated.married             : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Status...Sex_male...divorce.seperated                       : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Status...Sex_male.married.widowed                           : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Status...Sex_male.single                                    : int  1 0 1 1 1 1 1 1 0 0 ...
##  $ Property.owned_building.society.savings.agreement..life.insurance: int  0 0 0 1 0 0 1 0 0 0 ...
##  $ Property.owned_car.or.other                                 : int  0 0 0 0 0 0 0 1 0 1 ...
##  $ Property.owned_real.estate                                  : int  1 1 1 0 0 0 0 0 1 0 ...
##  $ Property.owned_unknown...no.property                        : int  0 0 0 0 1 1 0 0 0 0 ...
##  $ Purpose_business                                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Purpose_car.new.                                            : int  0 0 0 0 1 0 0 0 0 1 ...
##  $ Purpose_car.used.                                           : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Purpose_domestic.appliance                                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Purpose_education                                           : int  0 0 1 0 0 1 0 0 0 0 ...
##  $ Purpose_furniture.equipment                                 : int  0 0 0 1 0 0 1 0 0 0 ...
##  $ Purpose_others                                              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Purpose_radio.tv                                            : int  1 1 0 0 0 0 0 0 1 0 ...
```

```
##  $ Purpose_repairs                                           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Purpose_retraining                                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Other.Installemnt.plans_bank                              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Other.Installemnt.plans_none                              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Other.Installemnt.plans_stores                            : int  0 0 0 0 0 0 0 0 0 0 ...
```

```r
#4. Next we wanted to get summary statistics on certain continous variables:

summary(german$Current.Ac.status)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   1.001   2.000   3.000
```

```r
summary(german$Duration_transformed)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -1.4302790 -0.7493400 -0.2386350  0.0000001  0.2720700  2.8074940
```

```r
summary(german$Credit.amt_transformed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.1435 -0.7118 -0.3427  0.0000  0.2969  3.0809
```

```r
summary(german$credit_history_transformed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    2.00    1.66    2.00    4.00
```

```r
summary(german$Emp.Tenure)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   2.000   2.384   4.000   4.000
```

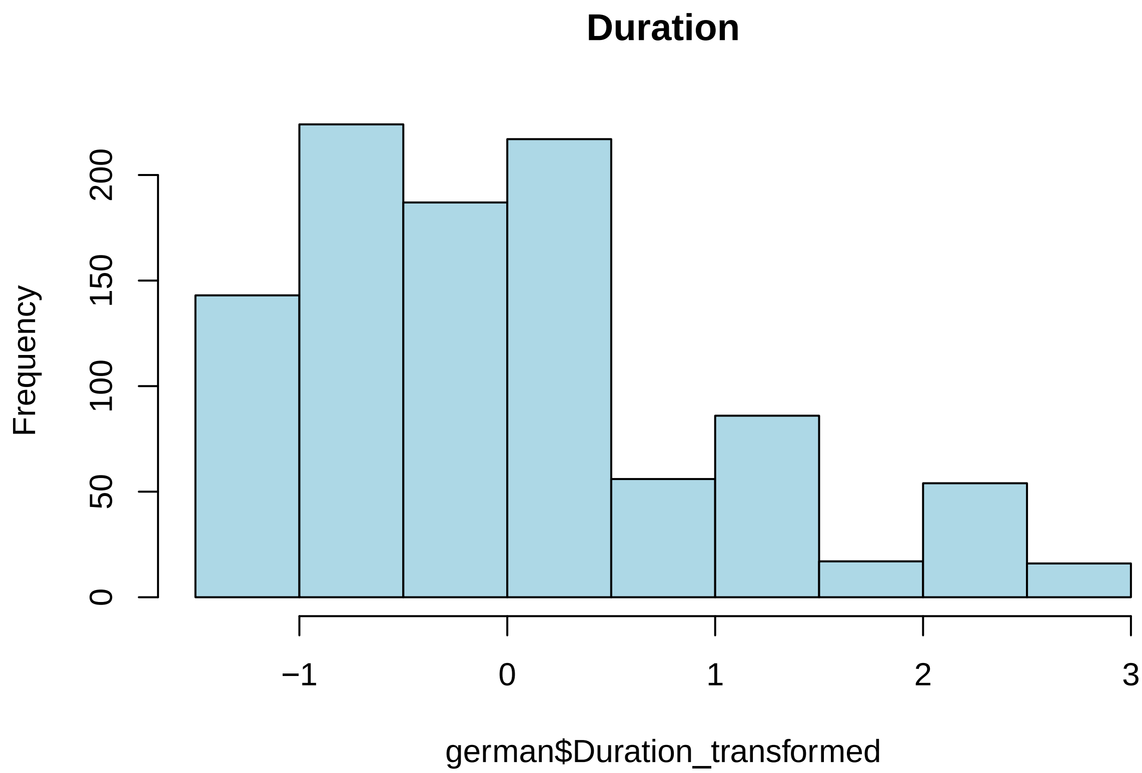
```r
summary(german$Age_transformed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.4670 -0.7560 -0.2228  0.0000  0.5770  2.8995
```

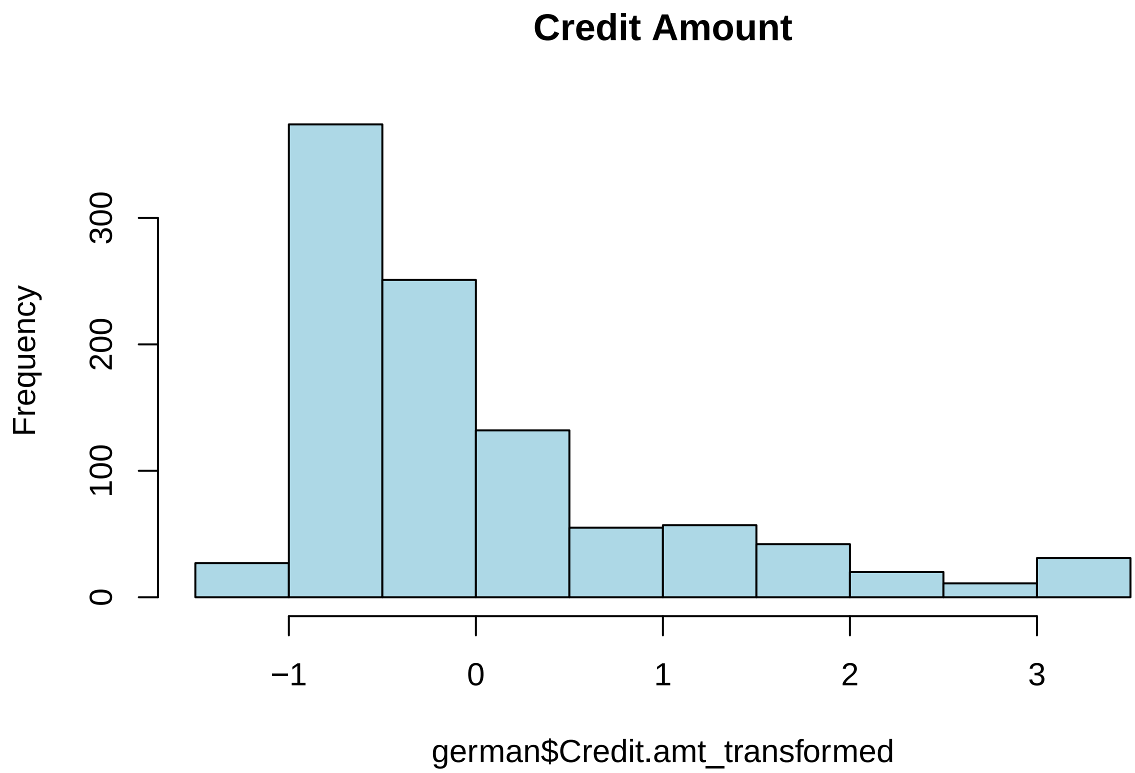
```r
summary(german$Installment.rate_transformed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.973   4.000   4.000
```

```r
#5. We then wanted to graphically plot them, since from summary statistics we understand that the data

hist(german$Duration_transformed, main = "Duration", col = "lightblue")
```

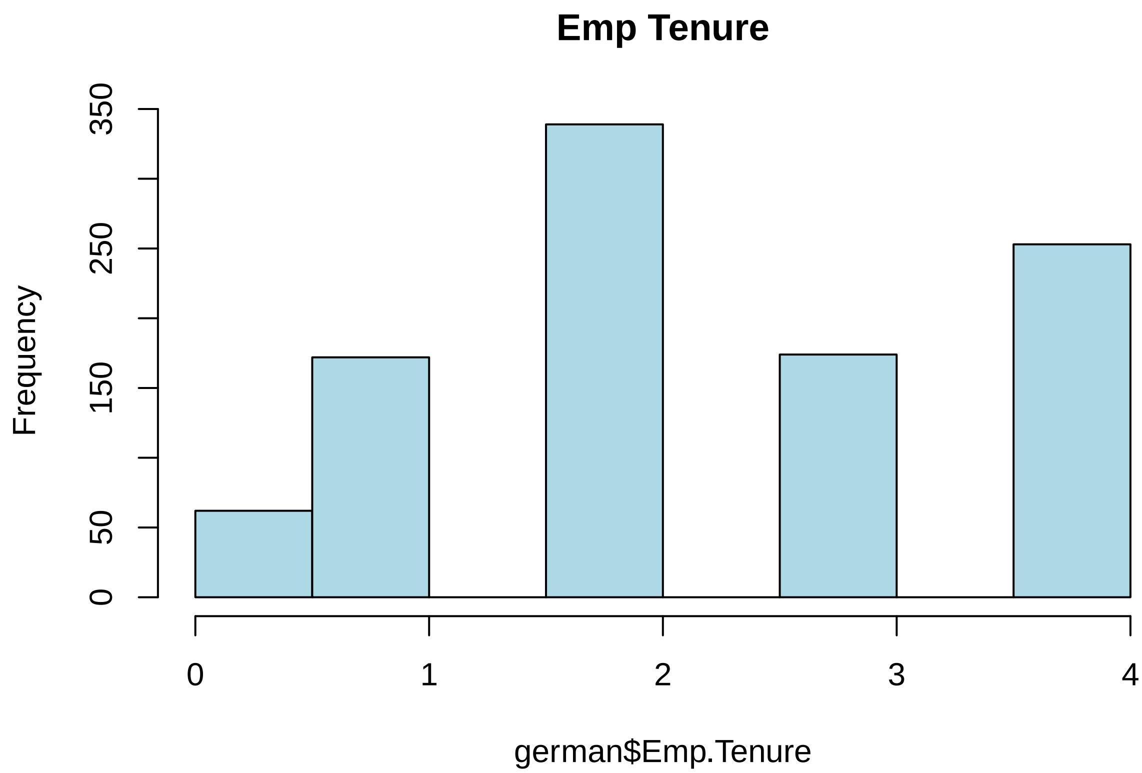



```
hist(german$Credit.amt_transformed, main = "Credit Amount", col = "lightblue")
```

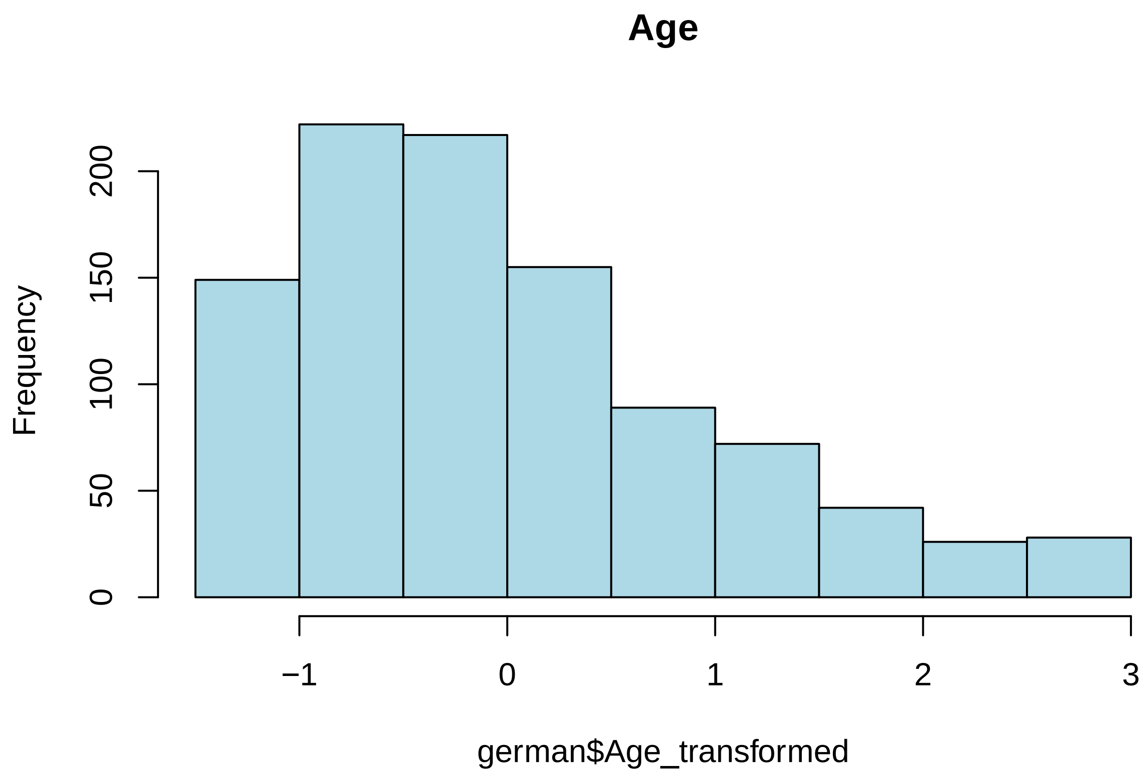



```r
hist(german$credit_history_transformed, main = "Credit History", col = "lightblue")
```


```
hist(german$Emp.Tenure, main = "Emp Tenure", col = "lightblue")
```



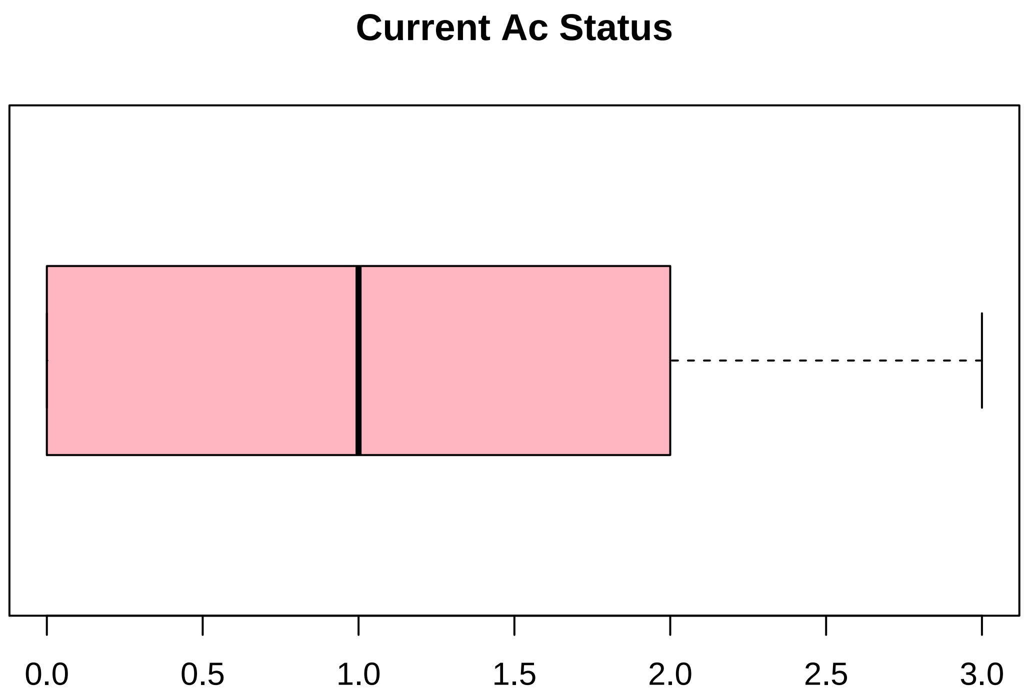

```r
hist(german$Age_transformed, main = "Age", col = "lightblue")
```



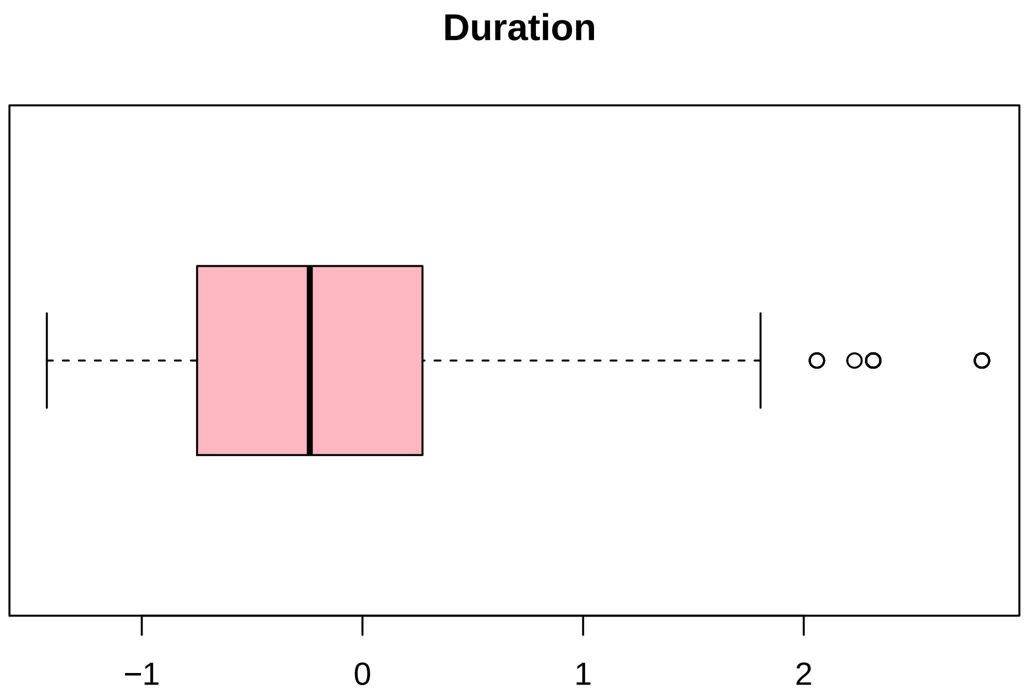

```
german<-german[,-1]
german$Credit.Risk<-as.factor(german$Credit.Risk)
#Mapping the classes as 0 and 1 for good and bad credit respectively. This was done becuase R GLM model
german$Credit.Risk<-mapvalues(german$Credit.Risk ,from =c("1","2"),
                        to = c("0","1"))

#checking the frequency of classes of target variable
table(german$Credit.Risk)
```

```
##
##   0   1
## 700 300
```

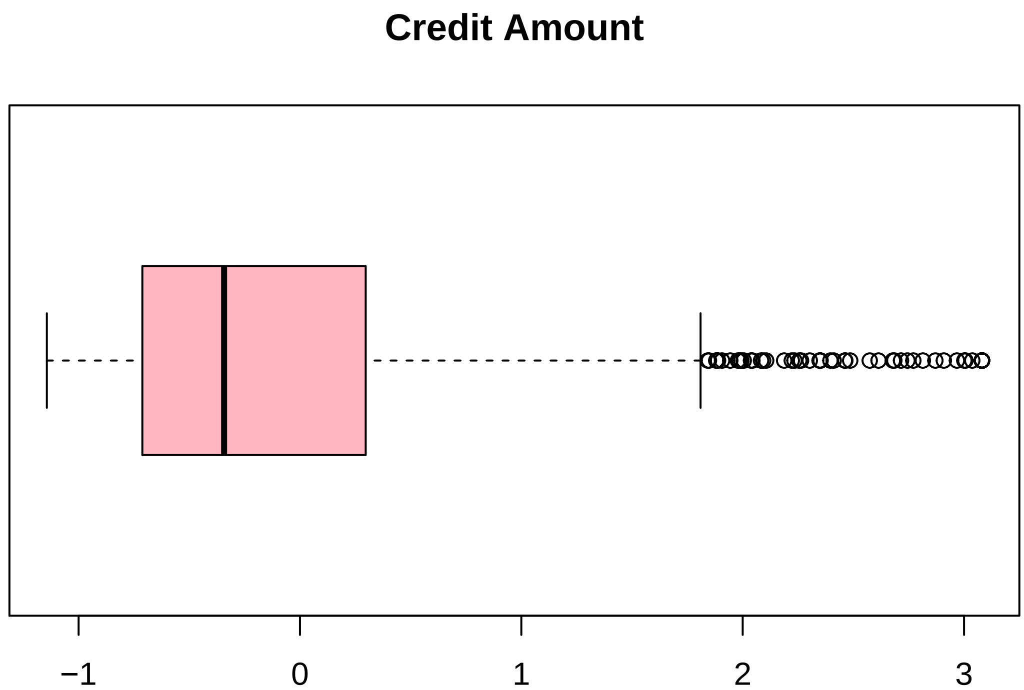
```
#6. We also wanted to plot box plots to understand the distribution of some continous variables better:

boxplot(german$Current.Ac.status, main = "Current Ac Status", horizontal = TRUE, col = "lightpink")
```



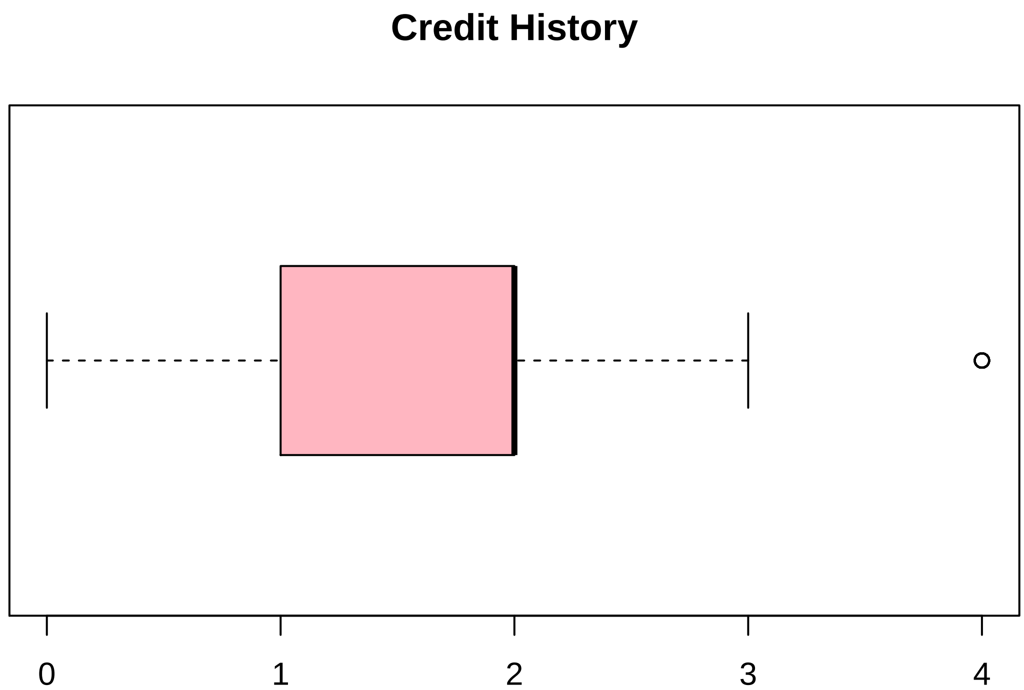

```
boxplot(german$Duration_transformed, main = " Duration", horizontal = TRUE, col = "lightpink")
```



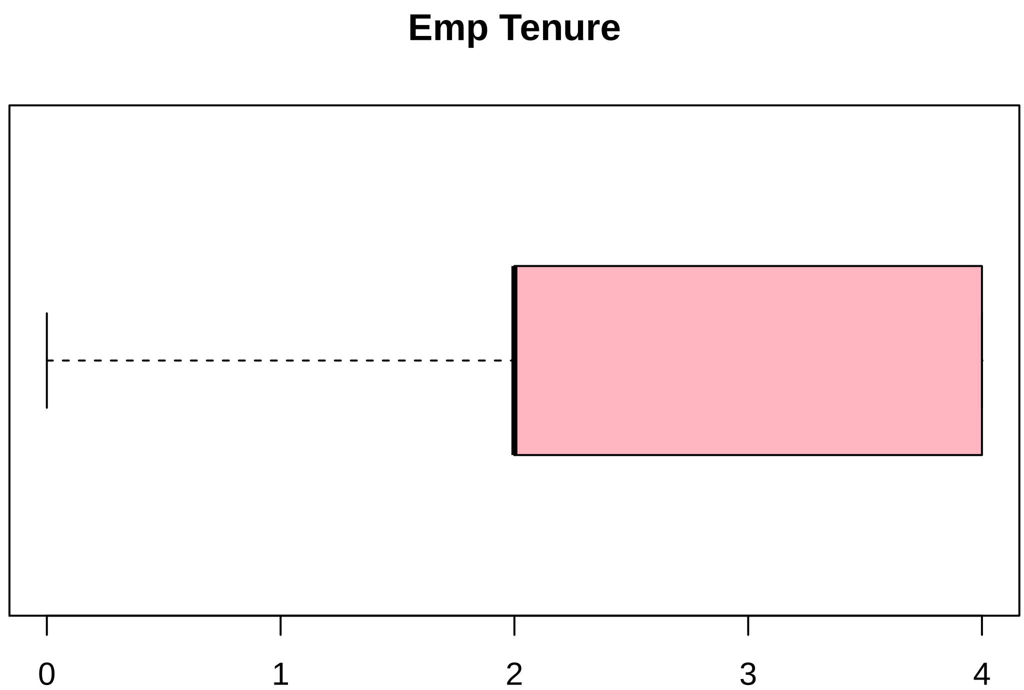

```
boxplot(german$Credit.amt_transformed, main = "Credit Amount", horizontal = TRUE, col = "lightpink")
```



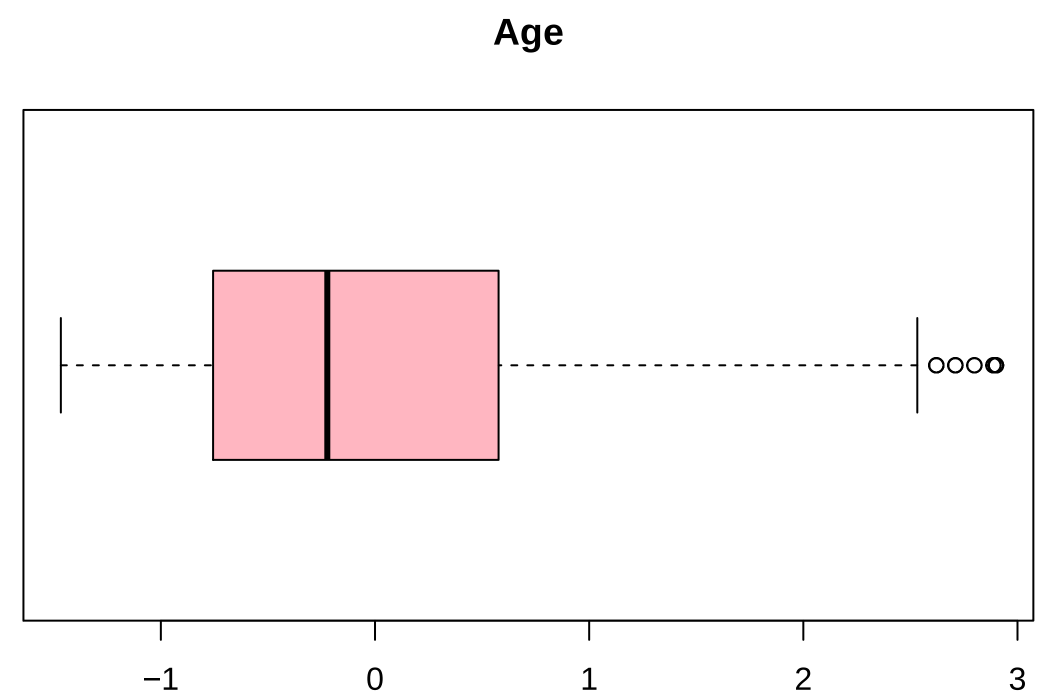

```
boxplot(german$credit_history_transformed, main = "Credit History", horizontal = TRUE, col = "lightpink
```



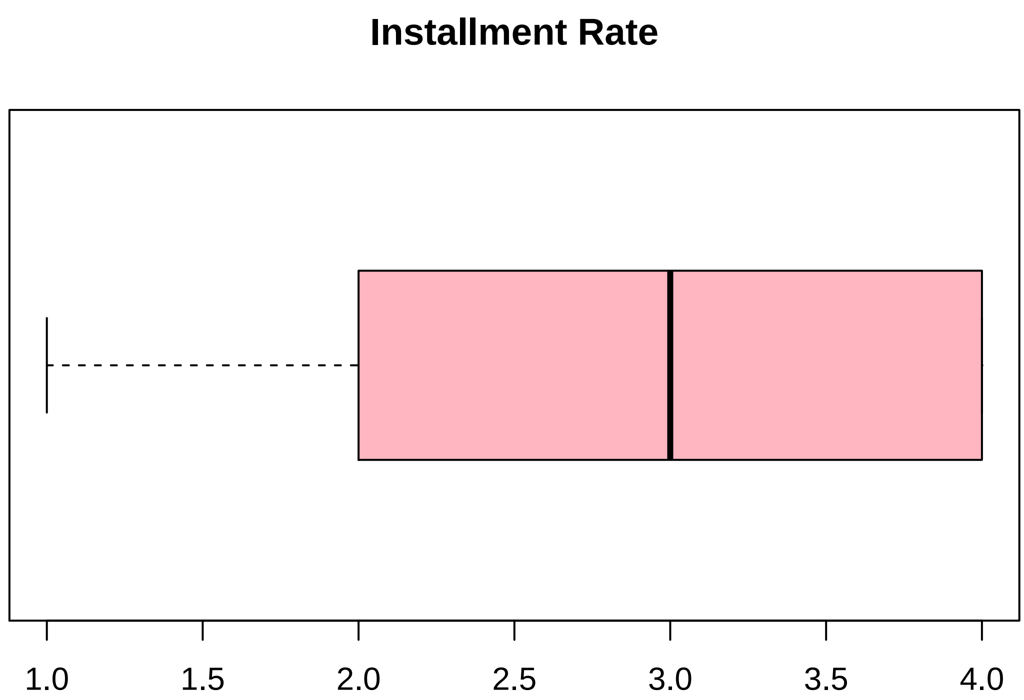

```r
boxplot(german$Emp.Tenure, main = "Emp Tenure", horizontal = TRUE, col = "lightpink")
```




```r
boxplot(german$Age_transformed, main = "Age", horizontal = TRUE, col = "lightpink")
```




```
boxplot(german$Installment.rate_transformed, main = "Installment Rate", horizontal = TRUE, col = "lightp
```




```
#7. Next, we wanted to understand the freq distribution of nominal variables and understand their propo

prop.table(table(german$Status...Sex_female...divorce.seperated.married))
```

```
##
##    0    1
## 0.69 0.31
```

```
prop.table(table(german$Status...Sex_male...divorce.seperated))
```

```
##
##    0    1
## 0.95 0.05
```

```
prop.table(table(german$Status...Sex_male.married.widowed))
```

```
##
##     0     1
## 0.908 0.092
```

```
prop.table(table(german$Status...Sex_male.single))
```

```
##
##     0     1
## 0.452 0.548
```

```r
prop.table(table(german$Property.owned_building.society.savings.agreement..life.insurance))
```

```
##
##     0     1
## 0.768 0.232
```

```r
prop.table(table(german$Property.owned_car.or.other))
```

```
##
##     0     1
## 0.668 0.332
```

```r
prop.table(table(german$Property.owned_real.estate))
```

```
##
##     0     1
## 0.718 0.282
```

```r
prop.table(table(german$Property.owned_unknown...no.property))
```

```
##
##     0     1
## 0.846 0.154
```

```r
prop.table(table(german$Purpose_business))
```

```
##
##     0     1
## 0.903 0.097
```

```r
prop.table(table(german$Purpose_car.new.))
```

```
##
##     0     1
## 0.766 0.234
```

```r
prop.table(table(german$Purpose_car.used.))
```

```
##
##     0     1
## 0.897 0.103
```

```r
prop.table(table(german$Purpose_domestic.appliance))
```

```
##
##     0     1
## 0.988 0.012
```

```r
prop.table(table(german$Purpose_education))
```

```
##
##    0    1
## 0.95 0.05
```

```r
prop.table(table(german$Purpose_furniture.equipment))
```

```
##
##     0     1
## 0.819 0.181
```

```r
prop.table(table(german$Purpose_others))
```

```
##
##     0     1
## 0.988 0.012
```

```r
prop.table(table(german$Purpose_radio.tv))
```

```
##
##    0    1
## 0.72 0.28
```

```r
prop.table(table(german$Purpose_repairs))
```

```
##
##     0     1
## 0.978 0.022
```

```r
prop.table(table(german$Purpose_retraining))
```

```
##
##     0     1
## 0.991 0.009
```

```r
#8. Lastly, we wanted to see the relationship between variables by ploting the scatter plot before we p

# install.packages("ggcorrplot")
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 3.6.3
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```r
corr <- round(cor(german[,2:14]), 1)
ggcorrplot(corr)
```


```r
#install.packages("caTools")
#install.packages("Rose")
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 3.6.3
```

```r
library(ROSE)
```

```
## Warning: package 'ROSE' was built under R version 3.6.3
```

```
## Loaded ROSE 0.0-3
```

```r
set.seed(123)
split_data = sample.split(german,SplitRatio = 0.8)
training = subset(german,split_data == TRUE)
test = subset(german,split_data == FALSE)

features<-setdiff(names(training),"Credit.Risk")
#Predictors created in SPSS for the model
print(features)
```

```
##  [1] "Installment.rate_transformed"
##  [2] "Residence.Tenure_transformed"
##  [3] "Existing.credit_transformed"
##  [4] "Dependents_transformed"
##  [5] "Duration_transformed"
##  [6] "Credit.amt_transformed"
##  [7] "Age_transformed"
##  [8] "Current.Ac.status"
##  [9] "SavingAc.Bonds"
## [10] "Emp.Tenure"
## [11] "Debtors.Guarantors"
## [12] "Housing"
## [13] "Job"
## [14] "Telephone"
## [15] "Foreign.Worker"
## [16] "credit_history_transformed"
## [17] "Status...Sex_female...divorce.seperated.married"
## [18] "Status...Sex_male...divorce.seperated"
## [19] "Status...Sex_male.married.widowed"
## [20] "Status...Sex_male.single"
## [21] "Property.owned_building.society.savings.agreement..life.insurance"
## [22] "Property.owned_car.or.other"
## [23] "Property.owned_real.estate"
## [24] "Property.owned_unknown...no.property"
## [25] "Purpose_business"
## [26] "Purpose_car.new."
## [27] "Purpose_car.used."
## [28] "Purpose_domestic.appliance"
## [29] "Purpose_education"
## [30] "Purpose_furniture.equipment"
## [31] "Purpose_others"
## [32] "Purpose_radio.tv"
## [33] "Purpose_repairs"
## [34] "Purpose_retraining"
## [35] "Other.Installemnt.plans_bank"
## [36] "Other.Installemnt.plans_none"
## [37] "Other.Installemnt.plans_stores"
```

```r
# Checking frequency of class of target variable in training data.
table(training$Credit.Risk)
```

```
##
##   0   1
## 556 233
```

```r
#Undersampling the training data to reduce imbalance between classes
data_balanced_under <- ovun.sample(Credit.Risk ~ ., data = training , method = "under", N = 466 , seed =
table(data_balanced_under$Credit.Risk)
```

```
##
##   0   1
## 233 233
```

```r
#Fit the logistic regression with training data set

logit_model<-glm(Credit.Risk~Installment.rate_transformed+Residence.Tenure_transformed+Existing.credit_
#checking the coefficent and the model description.
summary(logit_model)
```

```
##
## Call:
## glm(formula = Credit.Risk ~ Installment.rate_transformed + Residence.Tenure_transformed +
##     Existing.credit_transformed + Dependents_transformed + Duration_transformed +
##     Credit.amt_transformed + Age_transformed + Current.Ac.status +
##     SavingAc.Bonds + Emp.Tenure + Debtors.Guarantors + Housing +
##     Job + Telephone + Foreign.Worker + credit_history_transformed +
##     Status...Sex_female...divorce.seperated.married + Status...Sex_male...divorce.seperated +
##     Status...Sex_male.married.widowed + Status...Sex_male.single +
##     Property.owned_building.society.savings.agreement..life.insurance +
##     Property.owned_car.or.other + Property.owned_real.estate +
##     Property.owned_unknown...no.property + Purpose_business +
##     Purpose_car.new. + Purpose_car.used. + Purpose_domestic.appliance +
##     Purpose_education + Purpose_furniture.equipment + Purpose_others +
##     Purpose_radio.tv + Purpose_repairs + Purpose_retraining +
##     Other.Installemnt.plans_bank + Other.Installemnt.plans_none +
##     Other.Installemnt.plans_stores, family = binomial("logit"),
##     data = data_balanced_under)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.83285  -0.88584  -0.01363   0.90671   2.05192
##
## Coefficients: (4 not defined because of singularities)
##                                                                    Estimate
## (Intercept)                                                        -3.11618
## Installment.rate_transformed                                        0.35387
## Residence.Tenure_transformed                                        0.10205
## Existing.credit_transformed                                        -0.07400
## Dependents_transformed                                              1.21094
## Duration_transformed                                                0.30890
## Credit.amt_transformed                                              0.47162
## Age_transformed                                                    -0.33946
## Current.Ac.status                                                   0.26894
## SavingAc.Bonds                                                     -0.02060
## Emp.Tenure                                                         -0.16019
## Debtors.Guarantors                                                  0.07114
## Housing                                                            -0.36064
## Job                                                                 0.05110
## Telephoneyes                                                       -0.22993
## Foreign.Workeryes                                                   1.09499
## credit_history_transformed                                          0.42456
## Status...Sex_female...divorce.seperated.married                     0.76105
## Status...Sex_male...divorce.seperated                               0.89487
## Status...Sex_male.married.widowed                                   0.78706
## Status...Sex_male.single                                                 NA
## Property.owned_building.society.savings.agreement..life.insurance -0.60678
```

```
## Property.owned_car.or.other                                    -0.65526
## Property.owned_real.estate                                     -0.96074
## Property.owned_unknown...no.property                                 NA
## Purpose_business                                                1.17677
## Purpose_car.new.                                                1.55047
## Purpose_car.used.                                              -0.25170
## Purpose_domestic.appliance                                      0.96357
## Purpose_education                                               1.16526
## Purpose_furniture.equipment                                     0.87998
## Purpose_others                                                  0.41486
## Purpose_radio.tv                                                0.51091
## Purpose_repairs                                                 1.12389
## Purpose_retraining                                                   NA
## Other.Installemnt.plans_bank                                   -1.10081
## Other.Installemnt.plans_none                                   -1.73366
## Other.Installemnt.plans_stores                                       NA
##                                                                Std. Error
## (Intercept)                                                       1.94237
## Installment.rate_transformed                                      0.11737
## Residence.Tenure_transformed                                      0.11357
## Existing.credit_transformed                                       0.21888
## Dependents_transformed                                            0.36990
## Duration_transformed                                              0.14800
## Credit.amt_transformed                                            0.16725
## Age_transformed                                                   0.13929
## Current.Ac.status                                                 0.11955
## SavingAc.Bonds                                                    0.11980
## Emp.Tenure                                                        0.10058
## Debtors.Guarantors                                                0.25422
## Housing                                                           0.26578
## Job                                                               0.19258
## Telephoneyes                                                      0.26664
## Foreign.Workeryes                                                 0.71462
## credit_history_transformed                                        0.13460
## Status...Sex_female...divorce.seperated.married                   0.27638
## Status...Sex_male...divorce.seperated                             0.45759
## Status...Sex_male.married.widowed                                 0.40481
## Status...Sex_male.single                                               NA
## Property.owned_building.society.savings.agreement..life.insurance    0.44907
## Property.owned_car.or.other                                       0.43377
## Property.owned_real.estate                                        0.46769
## Property.owned_unknown...no.property                                   NA
## Purpose_business                                                  1.33821
## Purpose_car.new.                                                  1.29561
## Purpose_car.used.                                                 1.35831
## Purpose_domestic.appliance                                        1.54001
## Purpose_education                                                 1.35024
## Purpose_furniture.equipment                                       1.30417
## Purpose_others                                                    1.62573
## Purpose_radio.tv                                                  1.29135
## Purpose_repairs                                                   1.41445
## Purpose_retraining                                                     NA
## Other.Installemnt.plans_bank                                      0.69269
## Other.Installemnt.plans_none                                      0.65138
```

```
## Other.Installemnt.plans_stores                                             NA
##                                                                       z value
## (Intercept)                                                            -1.604
## Installment.rate_transformed                                            3.015
## Residence.Tenure_transformed                                            0.899
## Existing.credit_transformed                                            -0.338
## Dependents_transformed                                                  3.274
## Duration_transformed                                                    2.087
## Credit.amt_transformed                                                  2.820
## Age_transformed                                                        -2.437
## Current.Ac.status                                                       2.250
## SavingAc.Bonds                                                         -0.172
## Emp.Tenure                                                             -1.593
## Debtors.Guarantors                                                      0.280
## Housing                                                                -1.357
## Job                                                                     0.265
## Telephoneyes                                                           -0.862
## Foreign.Workeryes                                                       1.532
## credit_history_transformed                                              3.154
## Status...Sex_female...divorce.seperated.married                         2.754
## Status...Sex_male...divorce.seperated                                   1.956
## Status...Sex_male.married.widowed                                       1.944
## Status...Sex_male.single                                                   NA
## Property.owned_building.society.savings.agreement..life.insurance  -1.351
## Property.owned_car.or.other                                            -1.511
## Property.owned_real.estate                                             -2.054
## Property.owned_unknown...no.property                                       NA
## Purpose_business                                                        0.879
## Purpose_car.new.                                                        1.197
## Purpose_car.used.                                                      -0.185
## Purpose_domestic.appliance                                              0.626
## Purpose_education                                                       0.863
## Purpose_furniture.equipment                                             0.675
## Purpose_others                                                          0.255
## Purpose_radio.tv                                                        0.396
## Purpose_repairs                                                         0.795
## Purpose_retraining                                                         NA
## Other.Installemnt.plans_bank                                           -1.589
## Other.Installemnt.plans_none                                           -2.662
## Other.Installemnt.plans_stores                                             NA
##                                                                      Pr(>|z|)
## (Intercept)                                                           0.10864
## Installment.rate_transformed                                          0.00257 **
## Residence.Tenure_transformed                                          0.36887
## Existing.credit_transformed                                           0.73529
## Dependents_transformed                                                0.00106 **
## Duration_transformed                                                  0.03688 *
## Credit.amt_transformed                                                0.00481 **
## Age_transformed                                                       0.01481 *
## Current.Ac.status                                                     0.02447 *
## SavingAc.Bonds                                                        0.86350
## Emp.Tenure                                                            0.11122
## Debtors.Guarantors                                                    0.77960
## Housing                                                               0.17480
```

```
## Job                                                                 0.79075
## Telephoneyes                                                        0.38850
## Foreign.Workeryes                                                   0.12546
## credit_history_transformed                                          0.00161 **
## Status...Sex_female...divorce.seperated.married                     0.00589 **
## Status...Sex_male...divorce.seperated                               0.05051 .
## Status...Sex_male.married.widowed                                   0.05186 .
## Status...Sex_male.single                                                 NA
## Property.owned_building.society.savings.agreement..life.insurance  0.17663
## Property.owned_car.or.other                                         0.13089
## Property.owned_real.estate                                          0.03995 *
## Property.owned_unknown...no.property                                     NA
## Purpose_business                                                    0.37921
## Purpose_car.new.                                                    0.23142
## Purpose_car.used.                                                   0.85299
## Purpose_domestic.appliance                                          0.53152
## Purpose_education                                                   0.38814
## Purpose_furniture.equipment                                         0.49984
## Purpose_others                                                      0.79858
## Purpose_radio.tv                                                    0.69237
## Purpose_repairs                                                     0.42686
## Purpose_retraining                                                       NA
## Other.Installemnt.plans_bank                                        0.11202
## Other.Installemnt.plans_none                                        0.00778 **
## Other.Installemnt.plans_stores                                           NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 646.01  on 465  degrees of freedom
## Residual deviance: 510.73  on 432  degrees of freedom
## AIC: 578.73
##
## Number of Fisher Scoring iterations: 4
```

```
#Predicting over the balanced data based on above model
probability_model = predict(logit_model,type = 'response',newdata = data_balanced_under[-1])
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
prediction_y_train = ifelse(probability_model > 0.5,1,0)

#Predicting test result
probability_model = predict(logit_model,type = 'response',newdata = test[-1])
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
prediction_y_test = ifelse(probability_model > 0.5,1,0)
```

```r
confusion_matrix_training = table(data_balanced_under[,1], prediction_y_train)
confusion_matrix_training
```

```
##    prediction_y_train
##       0   1
##   0 165  68
##   1  69 164
```

```r
confusion_matrix_testing = table(test[,1], prediction_y_test)
confusion_matrix_testing
```

```
##    prediction_y_test
##      0  1
##   0 94 50
##   1 17 50
```